Impact of Outliers on K-means versus K-medoids

...

Online Encyclopedia of Integer Sequences (OEIS)

Github repository link: https://github.com/HeTalksInMaths/UnsupMLProj-OEIS

Project Overview

Attempt quantitative analysis on K-medoids being less susceptible to outliers than K-means.

Cluster first ten sequence values of sequences from OEIS.

Paired with and without outliers datasets to overcome no labels.

Use case - help determine if multiple models required for some downstream task.

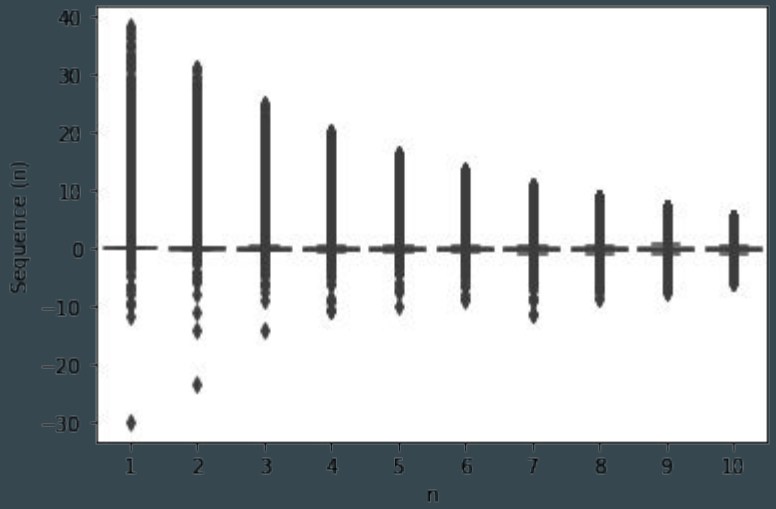
Imbalanced data sets - consider various metrics beyond accuracy.

Data

Online Encyclopedia of Integer Sequences (OEIS), 350,000 + seqs.

Curated with outliers, nearly 190,000 seqs.

Curated without outliers (Z-score between -3 and 3) , 175,000 + seqs.



Models

Elbow method, K = 3.

Paired K-means with and without outliers predicting one another.

Paired K-medoids with and without outliers predicting one another.

Sampling for K-medoids due to being memory intensive.

Results

Metric	K-means w/ outliers	K-means wo/ outliers	K-medoids w/ outliers	K-medoids wo/ outliers
Accuracy	0.864	0.841	0.899	0.904
F_1 Weighted	0.802	0.897	0.893	0.907

Metrics based on paired datasets.

K = 3 with all figures to 3 s.f.

Key Takeaways (for OEIS)

Given imbalanced clusters, weighted F_1 more appropriate metric.

As expected K-means with outliers performed worst when inferring K-means without outliers cluster labels.

Slight improvement for K-medoids without outliers but less variation.